Integrated Gradient Correlation:
a Dataset-wise Attribution Method

Pierre Lelièvre*
Department of Psychology
National Taiwan University
`contact@plelievre.com`

Chien-Chung Chen
Department of Psychology
National Taiwan University
`c3chen@ntu.edu.tw`

Abstract

Attribution methods are primarily designed to study the distribution of input component contributions to individual model predictions. However, some research applications require a summary of attribution patterns across the entire dataset to facilitate the interpretability of the scrutinized models. In this paper, we present a new method called Integrated Gradient Correlation (IGC) that relates dataset-wise attributions to a model prediction score and enables region-specific analysis by a direct summation over associated components. We demonstrate our method on scalar predictions with the study of image feature representation in the brain from fMRI neural signals and the estimation of neural population receptive fields (NSD dataset), as well as on categorical predictions with the investigation of handwritten digit recognition (MNIST dataset). The resulting IGC attributions show selective patterns, revealing underlying model strategies coherent with their respective objectives.

1   Motivation

Existing attribution methods study the distribution of input component contributions to individual model predictions. Here, we investigate the problem of summarizing these attributions at a dataset level. We particularly search for a method fulfilling the interpretability requirements of some modeling scenarios, e.g. *Where a specific image feature is represented in the visual cortex of the human brain?*, *What is the receptive field of a population of neurons?*, or *What is the model recognition strategy of handwritten digits?* (see Section.4 for details). These examples have in common that, beyond the need for models making accurate predictions, they call for the exposition of which input components are responsible for achieving the overall task, and not specific entries.

The implicit aspect of these scenarios is thus a *stable* localization of input information across the dataset. For instance, with fMRI (functional magnetic resonance imaging) signals, each input voxel/vertex represents an invariant brain area, covering a population of neurons. On the other hand, in the context of image content classification, a dataset-wise attribution method would not be pertinent: e.g. a *cat* can appear anywhere in an image while still accurately triggering the *cat* class.

Some research fields, such as neuroscience, heavily rely on linear or multilinear models to straightforwardly associate prediction scores with input components. Modelers overcome inherent limitations with *linearized* data, based on hand-crafted nonlinear preprocessings. Although legitimate, such an approach is unlikely to produce perfect linear mappings. Only deep models could exploit nonlinear and multiscale interactions without *a priori* data domain adaptation. However, the lack of dataset-wise summary techniques leaves deep models with an interpretability issue.

---

*git, website

Our new framework is therefore designed to be easily integrated in research activities and transparently used in place of linear regression analysis. To do so, it must fulfill requirements expressed in Naselaris et al. [20] with the following series of questions: "Does an input region of interest (ROI), contain information about some specific set of output features? Are there specific ROIs that contain relatively more information about a specific set of features? Are there specific features that are preferentially represented by a single ROI?" We then identify three main specifications: 1) a dataset-wise attribution method must permit a flexible definition of ROIs; 2) relative ROI attribution levels must allow pertinent comparisons; 3) dataset-wise attributions must enable comparisons between different features, and implicitly for identical features, but through different models.

1.1 General modelling context

We define a model as a function $f : \mathbb{R}^m \to \mathbb{R}$, predicting a random variable $\mathrm{y} \in \mathbb{R}$ from a random vector $\mathbf{x} \in \mathbb{R}^m$. This model is trained with n i.i.d pairs of a dataset $\mathcal{D}$, so that $(\mathbf{x}, \mathrm{y}) \sim \mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. f is also supposed to be a deep network optimized via gradient back-propagation, so that f is fully differentiable.[1]

2 Attribution methods for individual predictions

Available techniques for individual predictions arose from different research domains. For instance, the Shapley Value [2, 24], originated from cooperative game theory, was designed to address cost/gain sharing problems between players w.r.t their contributions to the outcome. Modern variants [17, 27] try to reduce the inherent computational cost. However, a method like Baseline Shapley (BS) still requires successive activations of each input component to measure associated contribution. In addition, this operation must be repeated several times with different random sequences of activated components. Such an approach is therefore not suited for large input dimensionalities.

More recently, image content recognition models prompted the development of methods to visualize pixel arrangements responsible for a classification decision. Relying on neural networks, these models are differentiable and led to techniques exploiting gradient back-propagation to reduce the computational cost. Nevertheless, naive gradient computations encounter issues with activation functions such as ReLU, as it back-propagates zero gradients for negative inputs. Consequently, methods like GradCAM [5, 23], LIME [22], DeepLift [25], or LRP [4] decided to operate architectural tweaks in models to avoid attribution shortcomings.

Another alternative to prevent spurious gradients is *Integrated Gradients* (IG) [28]. The idea is to aggregate the gradients of linearly interpolated inputs, from a baseline to the input under scrutiny (see Eq.9 for details). This way, even if some input components receive zero gradients, their comparison with possibly non-zero baseline gradients reveals more correct contributions. In addition, IG can be thought as a path method like BS. However, IG is not limited to the evaluation of all edges of a hypercube of m-dimension. IG can follow the shortest line from the baseline to the input, with all component attributions updated at once for few forward/backward passes through the model. IG is therefore quicker to compute than BS by several folds. Although popular, IG still exhibits unresolved issues, such as the optimality of selected paths [11, 12] or the choice of the most appropriate baseline[16, 29]. On this particular point, we favor the random baseline approach for its genericity (see below).

2.1 Definition

We define an attribution method as a function $g_{i,j}$ providing a contribution value for the $j \in [1, m]$ component of the input $i \in [1, n]$. This function takes for arguments a model f and an input random vector $\mathbf{x}$. Depending upon the method, an optional baseline $\bar{\boldsymbol{x}} \in \mathbb{R}^m$ may be required. $\bar{\boldsymbol{x}}$ is similar to $\mathbf{x}$ in terms of dimensionality, but it can be arbitrarily set to any value. For instance, $\bar{\boldsymbol{x}} = \mathbf{0}$ can simulate the least informational input in many cases, e.g. a black image.

However, defining a relevant baseline is difficult in some scenarios. For example, the activity of the brain in a resting state is rarely flat to zero. In such situations, we can turn the baseline into a random

---

[1]Actually, f has to be differentiable *almost everywhere*. See Sundararajan et al. [28] for more details.

vector $\bar{\mathbf{x}}$ sampled from the dataset $\mathcal{D}$ itself, and further consolidate attributions by taking the expected value of the different results obtained with multiple baselines. Then, we write:

$$g_{i,j}(f, \mathbf{x}, \bar{\mathbf{x}}) = \mathbb{E}_{\bar{\mathbf{x}} \sim \mathcal{D}}[g_{i,j}(f, \mathbf{x}, \bar{\mathbf{x}}_k)] \tag{1}$$

As random baseline methods are generalizations of single and non-baseline ones, we assume this variant in the following manuscript.

2.2 Axioms

With a careful axiomatization, Sundararajan et al. [28][27] clarified the categorization of attribution methods by revealing individual benefits and shortcomings. Among investigated properties, two of them are particularly important in the description of our method (see Section.3), and we describe them briefly.

Completeness   In a cost/gain sharing context, *completeness*, also denominated *efficiency*, refers to the intuitive idea that the sum of all contributions must be equal to the cost/gain under scrutiny. More generally, it means that the sum of all component attributions must reflect the sign and the magnitude of model predictions, so that:

$$g_i(f, \mathbf{x}, \bar{\boldsymbol{x}}) = \sum_{j=1}^{m} g_{i,j}(f, \mathbf{x}, \bar{\boldsymbol{x}}) = f(\mathbf{x}) - f(\bar{\boldsymbol{x}}) \tag{2}$$

In the random baseline scenario, it becomes:

$$g_i(f, \mathbf{x}, \bar{\mathbf{x}}) = \sum_{j=1}^{m} g_{i,j}(f, \mathbf{x}, \bar{\mathbf{x}}) = f(\mathbf{x}) - \mu_{f(\bar{\mathbf{x}})} \tag{3}$$

with $\mu_{f(\bar{\mathbf{x}})}$ the mean of baseline model predictions (see Appendix.A.1 for a complete development).

Implementation invariance   *Implementation invariance* refers to the fact that two functionally equivalent models (i.e. exhibiting similar input/output behaviors), but with different architectures (e.g. number and type of layers, or activation functions), should have identical attributions. For instance, attribution methods that require low-level modifications in models depending on selected activation functions, such as DeepLIFT [25], break this axiom. Moreover, beyond the inherent difficulty to implement this type of method, the status of implied architectural changes becomes unclear when attempting to explain the default behaviors of models.

3   Dataset-wise attribution methods

We consider dataset-wise attribution methods as an extension to classical methods for individual predictions. As a result, it can be formally written as a function h_j (w.r.t. each $j \in [1, m]$ component), taking as argument an attribution method $g_{i,j}$, a model f, an input $\mathbf{x}$, and a baseline $\bar{\mathbf{x}}$.

Additive property.   In order to fulfill first two interpretability requirements expressed earlier, $h_j(g_{i,j}, f, \mathbf{x}, \bar{\mathbf{x}})$ must be easily aggregated over ROIs. One possibility is to consider attribution values as a distribution over input components located in the ROI under scrutiny. However, the choice of an appropriate global value to report is difficult. For instance, in the context of brain data, we could think that information useful for a task is processed only by few neurons so that the maximum value would be more appropriate than the mean value. As a counterpart, such values would become very sensitive to outliers of noisy acquisitions. Therefore, we decide to enforce a stronger constraint. Summaries of attributions over ROIs must be obtained by a simple summation. So, for any ROI $\mathcal{R}$, we want:

$$h_{\mathcal{R}}(g_{i,j}, f, \mathbf{x}, \bar{\mathbf{x}}) = \sum_{j \in \mathcal{R}} h_j(g_{i,j}, f, \mathbf{x}, \bar{\mathbf{x}}) \tag{4}$$

Completeness to a model prediction score. The third requirement is to be able to compare dataset-wise attributions between different features, as well as identical features through different models. Consequently, dataset-wise attribution methods must first inherit the *implementation invariance* property from its supporting attribution method for individual predictions $g_{i,j}$.[2] Secondly, the total attribution over input components must be related to a prediction score h of the model. The *true* output y is then required as a parameter for h (and h_j). By extension to Eq.4, we write:

$$h(g_{i,j}, f, \mathbf{x}, \bar{\mathbf{x}}, \mathrm{y}) = \sum_{j=1}^{m} h_j(g_{i,j}, f, \mathbf{x}, \bar{\mathbf{x}}, \mathrm{y}) \tag{5}$$

3.1 Correlation-based dataset-wise attribution methods

We choose to use the correlation between predicted outputs $f(\mathbf{x})$ and *true* outputs y as the model prediction score h. So, by definition:

$$h = \rho_{f(\mathbf{x}),\mathrm{y}} = \frac{\mathbb{E}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}}[(f(\boldsymbol{x}_i) - \mu_{f(\mathbf{x})})(y_i - \mu_\mathrm{y})]}{\sigma_{f(\mathbf{x})}\sigma_\mathrm{y}} \tag{6}$$

with $\mu_{f(\mathbf{x})}$, μ_y, $\sigma_{f(\mathbf{x})}$, and σ_y respective means and standard deviations of $f(\mathbf{x})$ and y. To further develop this expression and introduce $g_{i,j}$, we reduce the possible spectrum of attribution methods for individual predictions to ones satisfying *completeness* (see Eq.3). Then:

$$
\begin{aligned}
h &= \frac{\mathbb{E}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}}[(\sum_{j=1}^{m} g_{i,j}(f, \boldsymbol{x}_i, \bar{\mathbf{x}}) + \mu_{f(\bar{\mathbf{x}})} - \mu_{f(\mathbf{x})})(y_i - \mu_\mathrm{y})]}{\sigma_{f(\mathbf{x})}\sigma_\mathrm{y}} \\
&= \frac{\mathbb{E}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}}[\sum_{j=1}^{m} g_{i,j}(f, \boldsymbol{x}_i, \bar{\mathbf{x}}) \times (y_i - \mu_\mathrm{y})]}{\sigma_{f(\mathbf{x})}\sigma_\mathrm{y}} + \frac{\mathbb{E}_{\mathrm{y}\sim\mathcal{D}}[(\mu_{f(\bar{\mathbf{x}})} - \mu_{f(\mathbf{x})})(y_i - \mu_\mathrm{y})]}{\sigma_{f(\mathbf{x})}\sigma_\mathrm{y}} \\
&= \sum_{j=1}^{m} \frac{\mathbb{E}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}}[g_{i,j}(f, \boldsymbol{x}_i, \bar{\mathbf{x}}) \times (y_i - \mu_\mathrm{y})]}{\sigma_{f(\mathbf{x})}\sigma_\mathrm{y}} + \frac{(\mu_{f(\bar{\mathbf{x}})} - \mu_{f(\mathbf{x})})(\cancel{\mathbb{E}_{\mathrm{y}\sim\mathcal{D}}[y_i] - \mu_\mathrm{y}})}{\cancel{\sigma_{f(\mathbf{x})}\sigma_\mathrm{y}}}
\end{aligned} \tag{7}
$$

Consequently, correlation-based dataset-wise attribution methods are defined as:

$$h_j(g_{i,j}, f, \mathbf{x}, \bar{\mathbf{x}}, \mathrm{y}) = \frac{1}{\sigma_{f(\mathbf{x})}\sigma_\mathrm{y}} \mathbb{E}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}}[g_{i,j}(f, \boldsymbol{x}_i, \bar{\mathbf{x}}) \times (y_i - \mu_\mathrm{y})] \tag{8}$$

3.2 Supporting attribution method for individual predictions

From the above development, supporting attribution method $g_{i,j}$ must satisfy *completeness* and *implementation invariance* axioms, described in Subsection.2.2. As a result, we have the choice between path methods like Baseline Shapley [6, 17, 27] and Integrated Gradients [28]. Each technique provides different attributions since they take different paths from baselines to inputs. However, our empirical results with BS do not show significant differences from IG (see Appendix.A.3). Then, because of BS expensive computational cost, we recommend the use of IG in the general case. Consequently, we call our dataset-wise attribution method: Integrated Gradient Correlation (IGC).

IG is originally defined with an integral, but Sundararajan et al. [28] propose to use in practice its Riemman approximation with few discrete s steps:

$$g_{i,j}(f, \mathbf{x}, \bar{\boldsymbol{x}}) \approx \frac{(\mathrm{x}_j - \bar{x}_j)}{s} \sum_{t=1}^{s} \frac{\partial f(\bar{\boldsymbol{x}} + \frac{t}{s}(\mathbf{x} - \bar{\boldsymbol{x}}))}{\partial \mathrm{x}_j} \tag{9}$$

---

[2]We empirically demonstrate the relative independence of our IGC results upon architectural changes in Appendix.A.2.
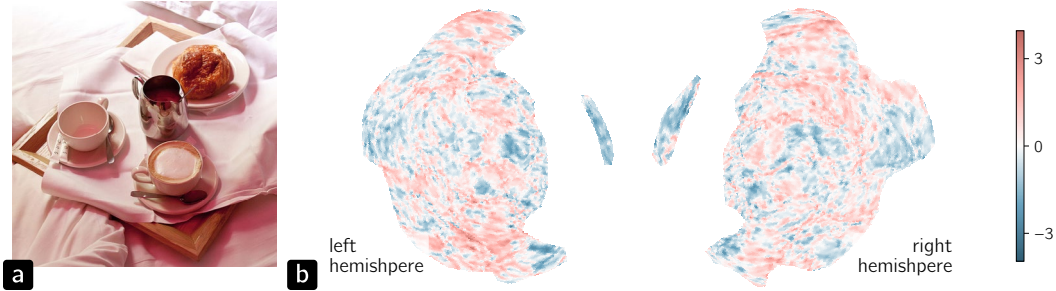
Figure 1: Panel **a** displays an image sampled from the NSD dataset, and panel **b** shows corresponding fMRI activation maps for left and right hemispheres. These surface-based fMRI data are projections from volumetric acquisition to an average brain space (fsaverage), limited to the visual cortex. Per vertex activations over the dataset are standardized per subjects.

3.3 IGC with categorical distributions

The present method is primarily designed for models predicting scalars. However, some modeling scenarios involve the prediction of random variables ruled by categorical distributions with K categories. Usually expressed as *one-hot* vectors, models are then actually predicting probabilities p_k for $k \in [1, K]$. A prediction score based on the correlation between p_k and corresponding *true one-hot* vector components is not the most appropriate, but it proved to be effective in practice (see Subsection.4.3). In addition, this setup allows the computation of dataset-wise attributions w.r.t. each class independently.

4 Applications

Integrated Gradient Correlation is applicable to a diversity of modeling scenarios. Here, we present a decoding model of fMRI data for investigating the representation of image statistics in the brain. We also introduce an encoding model, estimating the receptive field of a population of neurons. These two examples rely on scalar prediction, so we further explore a handwritten digit recognition model to demonstrate IGC on categorical predictions. For information, all presented IGC results use supporting IG values computed with 32 random baselines and 32 interpolation steps.

4.1 Representation of image statistics in the brain

With fMRI, the brain activity is measured through the proxy of the blood oxygenation level. This intermediary is relatively delayed from underlying electric neural activity but provides high-resolution activation maps. This technique is widely used in neuroscience.

The recent publicly available *Natural Scene Dataset* [1, 9] (NSD dataset) has been designed to be large enough to enable machine learning inquiries. It offers fMRI data for >70k distinct images, acquired during a long-term recognition task with 8 subjects over one year. Selected images come from the COCO dataset [15] (see Fig.1a). The fMRI data are initially volume-based, with brain hemispheres discretized as voxels, but they can also be projected on the surface of the brain gray matter, which is a convoluted layer where most activations happen. As result, such a surface is easier to map to an average brain morphology shared by all participants (the *fsaverage* template of *FreeSurfer* software) and enable inter-subject data aggregation. Here, we focus on left and right visual cortex, i.e. 2 graphs of nearly 20k vertices each (see Fig.1b).

For this first example, we build a model predicting two simple image statistics from fMRI data. Selected features are globally computed on gray-scaled images: luminance contrast (i.e. standard deviation) and *1/f slope* value. The first statistic is perhaps the most studied image variable in vision science whereas the second summarizes spatial frequency distribution and proved to be relevant to visual perception [8, 30, 31]. The slope refers to the decreasing intensities of higher spatial frequencies of natural images. Our model is a *vanilla* multilayer perceptron (6 layers) with batch-normalization [10] and Mish activation functions [19]. Despite the morphological mapping to an average brain, the functional behavior of each vertex may differ from one subject to another. Therefore, besides fMRI
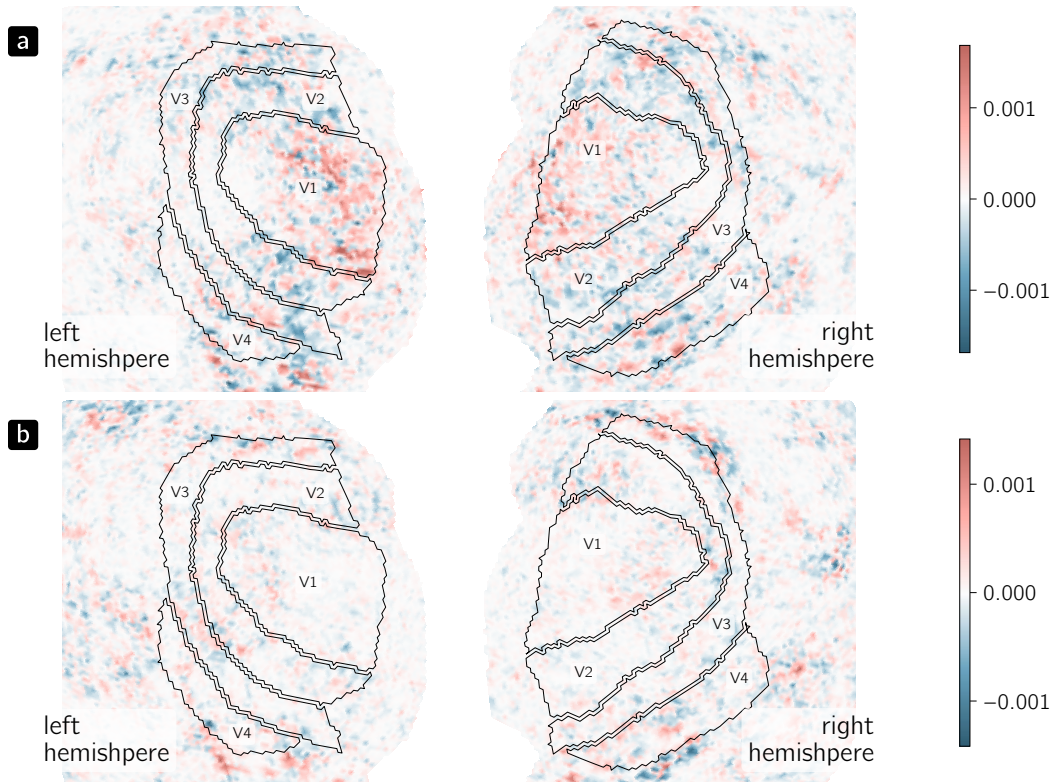
Figure 2: IGC maps associated with the prediction of image luminance contrast (panel **a**) and *1/f slope* (panel **b**) from fMRI data. IGC maps and outlines of early visual ROIs, correspond to the subject 1 of the NSD dataset.
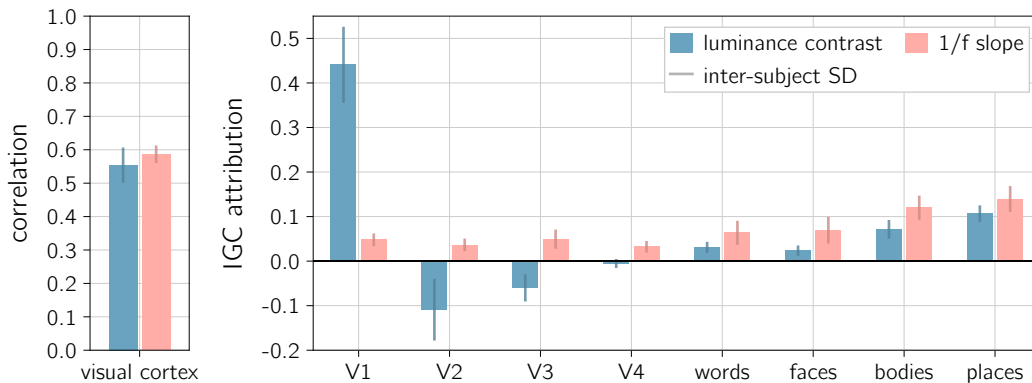

Figure 3: ROI-based summary of IGC attributions associated with the prediction of image luminance contrast and *1/f slope* from fMRI data. Error bars reflect inter-subject variability (1 SD).

data, we also provide a subject ID to the model. The trained network achieves a correlation of 0.56 for the luminance contrast and 0.59 for *1/f slope*.

Fig.2 shows resulting IGC maps for the first subject. In overlay, we display V1 to V4 ROIs that constitute the beginning of the visual pathway in the brain. Early convolutional blocks of deep image content recognition models, such as AlexNet [13] or VGG [26] share some arguable degree of similarity with these early visual ROIs [7, 21, 35]. So, for the luminance contrast in Fig.2a, we observe that V1 contributions are strongly positive (in red), while V2, V3 and V4 attributions appear negative (in blue). On the other hand, Fig.2b displays more diffuse patterns for *1/f slope*, that especially occur beyond V1 to V4 ROIs.
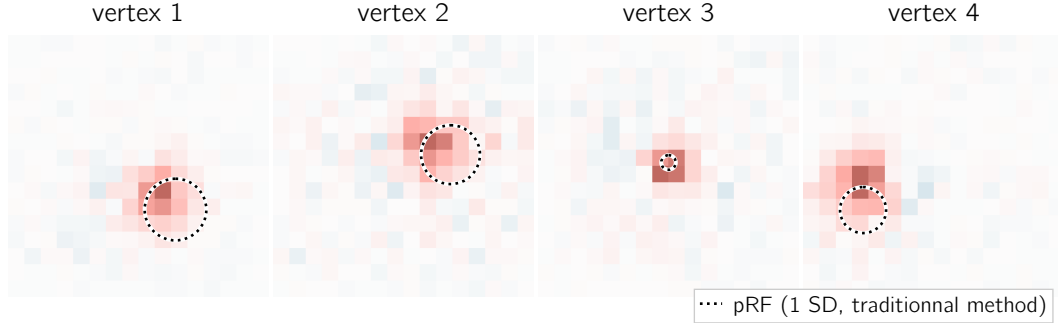
Figure 4: pRF estimation from IGC maps associated with the prediction of brain activity (fMRI data) from image stimuli. Selected vertices are located in V1 and correspond to the subject 1 of the NSD dataset. IGC maps are compared with the pRF computed with the traditional approach (dotted circles). For legibility, IGC maps have been downscaled by a local summation of pixels.

Due to the additive property of IGC attributions (see Eq.4), a more quantitative analysis is possible by a direct summation over ROIs. Then, Fig.3 confirms previously observed trends. Nonetheless, how to interpret negative attributions found in luminance contrast results? All IGC entries sum up to the overall model correlation performance, so we may think that excluding V2 and V3 ROIs from the model could produce higher performance. However, in this model we predict scalars. So, inputs that systematically negatively contribute to the prediction should rather be understood as an adjusting mechanism, balancing possibly overestimated scalar magnitude in other positive regions. Therefore, V2 and V3 probably serve as counterbalances to V1, which provides most of the luminance contrast estimation of entire images. The literature indicates that neurons with large receptive fields and associated with the peripheral vision are already present in V1, so the strategy revealed by our method seems plausible to reflect real neural mechanisms. Another aspect making us confident about IGC findings, is that higher-level brain areas dedicated to *bodies* and *places* appear more relevant to the task than *words* and *faces* ROIs.

Concerning *1/f slope*, IGC demonstrates a contrasting attribution distribution over the visual pathway. Fig.3 presents a more even use of all areas, with a preference for higher-level ROIs. This distinct strategy is in fact pertinent for a statistic summarizing spatial frequencies because the hierarchical processing of multiscale information is an inherent characteristic of our complete visual system.

4.2 Population receptive field

Neurons in early visual areas have been proven to be spatially specialized [3, 18, 32, 33, 34]. Hierarchically folded repetitions of the visual field in the brain is even what enables the definition of V1 to V4 ROIs. With a long history in neuroscience, this functional identification of brain regions has been addressed by specific methods that require dedicated fMRI acquisitions and traditional parameter optimizations.

In this second example, we propose to use our method on the NSD dataset to compute the population receptive field (pRF), or the summed responsible area in an image covered by a group of neurons. To this end, we build a simple model predicting fMRI data from images. It consists of five convolution layers, followed by two fully connected layers. We use batch-normalization and Mish activation functions. Images are downscaled to 128 pixels, and for similar reasons as the model described above, we concatenate a subject ID (encoded by one additional layer) to the input of the first linear layer. The model is trained to predict fMRI activations of all vertices, but for legibility, we limit our investigations to four representative ones, located in V1. The correlations between predictions and *true* activations are respectively 0.34, 0.45, 0.22, and 0.63.

Fig.4 shows that resulting IGC maps give a direct visualization of the pRF associated with each vertex. In addition, they are in coherence with *traditional* pRF provided by the NSD dataset (dotted circles).[3] They are not completely overlapped, but *traditional* pRF must not be considered as the ground truth. Their computation involves an arbitrary luminance definition, some local contrast

---

[3]pRF are usually modeled by 2-dimensional Gaussian distributions and reported with circles of radius 1 SD.
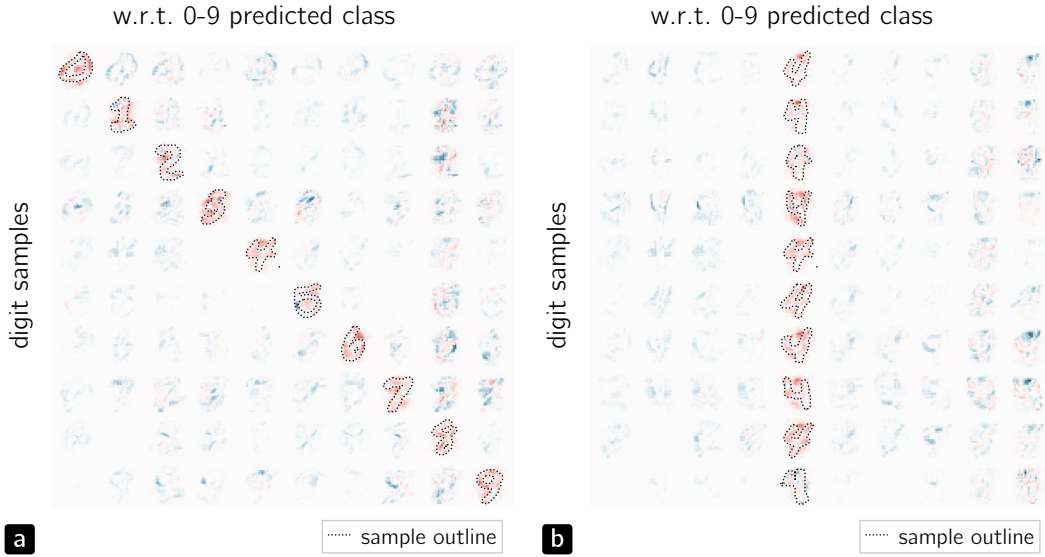
Figure 5: Integrated Gradients maps of MNIST samples w.r.t. all possible digit classes: for *0-9* samples in panel **a**, and ten different *4* samples in panel **b**.
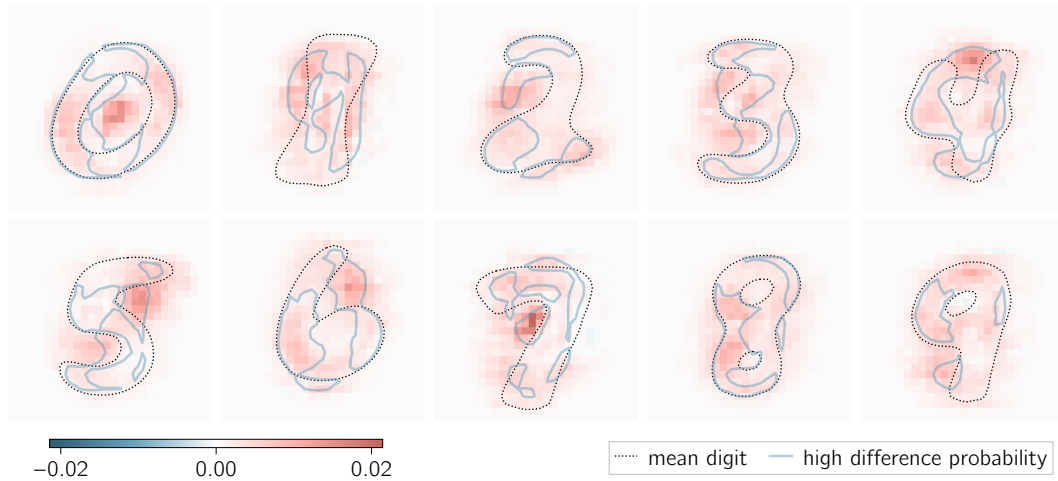

Figure 6: IGC maps for the MNIST dataset w.r.t. all possible digit classes. Dotted lines indicate mean digit contours (level is set to 80% of pixel values CDF). Blue solid lines show areas of pixels with a high probability to be different from other digit classes (level is set to 0.5).

approximations, and several uncertain optimization steps. In this sense, our method is more generic and straightforward. Nonetheless, attesting of a potential better accuracy of the IGC approach requires further investigations, left for future works.

4.3 Recognition strategy of handwritten digits

Finally, we illustrate our method with the prediction of a categorical distribution. Specifically, we investigate the recognition strategy of handwritten digits from the MNIST dataset [14]. Our model possesses two convolution layers, followed by five fully connected layers. It uses batch-normalization and Mish activation functions. The accuracy of the trained model is over 99%.

As described in Subsection.3.3, the model output probabilities p_k with $k \in [0, 9]$, and attributions are then computed per class. In Fig.5, we show Integrated Gradients of sampled digits w.r.t. all possible classes; for *0-9* samples in panel **a**, and 10 different *4* samples in panel **b**. These examples

show certain patterns, like a strong positive attribution in between the top branches of the *4*, but exact positions differ from one another. In addition, how to combine attributions (mostly negative) received for digits of other categories? This is exactly where the IGC summary over the whole dataset becomes useful.

In Fig.6, we remark that high IGC attributions are coherent with mean digit contours (dotted lines), but the critical locations are sometimes surprisingly outside digit bodies (e.g. in the center of *0*, or inside the angle of *7*). To better understand the logic behind it, we also plot with blue solid lines, the locations of pixels with a high probability to be different from other digit classes (level is set to 0.5). Many high IGC attributions are aligned with these areas (e.g. for *6* and *8*), but they are not limited to. It thus demonstrates that the recognition model integrates conditional inter-pixel mechanisms that go beyond a naive resolution of the task.

5 Conclusion

The main contribution of this paper is the definition of a dataset-wise attribution method, Integrated Gradient Correlation, that improves the interpretability of deep neural networks for research scenarios where the localization of input information is stable across the dataset. Resulting summarizing maps show selective attribution patterns, that reveal underlying model strategies coherent with their respective objectives.

A secondary contribution is that we introduced IGC as a particular case of dataset-wise attribution methods, generically defined by desirable characteristics, i.e. ROI attributions computed as the sum of associated components, and a total attribution related to a prediction score of the model under scrutiny.

In particular, our method uses correlation as a versatile prediction score, and Integrated Gradients as its supporting attribution method for individual predictions. IGC is therefore easy to implement, fast to compute, and generic enough to be applicable to a wide range of model architectures and data.

Acknowledgments and Disclosure of Funding

References

[1] Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022. ISSN 1097-6256, 1546-1726. doi:10.1038/s41593-021-00962-x.

[2] Robert J. Aumann and Lloyd Stowell Shapley. *Values of Non-Atomic Games*. Princeton University Press, Princeton, N.J, 1974. ISBN 978-0-691-08103-8.

[3] Noah C. Benson, Keith W. Jamison, Michael J. Arcaro, An T. Vu, Matthew F. Glasser, Timothy S. Coalson, David C. Van Essen, Essa Yacoub, Kamil Ugurbil, Jonathan Winawer, and Kendrick Kay. The Human Connectome Project 7 Tesla retinotopy dataset: Description and population receptive field analysis. *Journal of Vision*, 18(13):23, December 2018. ISSN 1534-7362. doi:10.1167/18.13.23.

[4] Alexander Binder, Grégoire Montavon, Sebastian Bach, Klaus-Robert Müller, and Wojciech Samek. Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers, April 2016. URL http://arxiv.org/abs/1604.00825.

[5] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, March 2018. doi:10.1109/WACV.2018.00097.

[6] Kedar Dhamdhere, Ashish Agarwal, and Mukund Sundararajan. The Shapley Taylor Interaction Index, February 2020. URL http://arxiv.org/abs/1902.05622.

[7] Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, May 2017. ISSN 10538119. doi:10.1016/j.neuroimage.2016.10.001.

[8] David J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379, December 1987. ISSN 1084-7529, 1520-8532. doi:10.1364/JOSAA.4.002379.

[9] Alessandro T Gifford, Benjamin Lahner, Sari Saba-Sadiya, Martina G Vilas, Aude Oliva, Kendrick Kay, Gemma Roig, and Radoslaw M Cichy. Algonauts Project 2023. 2023.

[10] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, March 2015. URL http://arxiv.org/abs/1502.03167.

[11] Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided Integrated Gradients: An Adaptive Path Method for Removing Noise, June 2021. URL http://arxiv.org/abs/2106.09788.

[12] Yuji Kawai, Kazuki Tachikawa, Jihoon Park, and Minoru Asada. Compensated Integrated Gradients for Reliable Explanation of Electroencephalogram Signal Classification. *Brain Sciences*, 12(7):849, June 2022. ISSN 2076-3425. doi:10.3390/brainsci12070849.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*, 25, January 2012. doi:10.1145/3065386.

[14] Yann LeCun, Corinna Cortes, and Christopher J.C. Burgess. The MNIST Database of Handwritten Digits, 1998. URL http://yann.lecun.com/exdb/mnist/.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context, February 2015. URL http://arxiv.org/abs/1405.0312.

[16] Shuyang Liu, Zixuan Chen, Ge Shi, Ji Wang, Changjie Fan, Yu Xiong, Runze Wu Yujing Hu, Ze Ji, and Yang Gao. Rethink Baseline of Integrated Gradients from the Perspective of Shapley Value, October 2023. URL http://arxiv.org/abs/2310.04821.

[17] Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions, November 2017. URL http://arxiv.org/abs/1705.07874.

[18] Wayne E Mackey, Jonathan Winawer, and Clayton E Curtis. Visual field map clusters in human frontoparietal cortex. *eLife*, 6:e22974, June 2017. ISSN 2050-084X. doi:10.7554/eLife.22974.

[19] Diganta Misra. Mish: A Self Regularized Non-Monotonic Activation Function, August 2020. URL http://arxiv.org/abs/1908.08681.

[20] Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, May 2011. ISSN 10538119. doi:10.1016/j.neuroimage.2010.07.073.

[21] Peter Neri. Deep networks may capture biological behavior for shallow, but not deep, empirical characterizations. *Neural Networks*, 152:244–266, August 2022. ISSN 08936080. doi:10.1016/j.neunet.2022.04.023.

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, August 2016. URL http://arxiv.org/abs/1602.04938.

[23] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv:1610.02391 [cs]*, October 2016. URL http://arxiv.org/abs/1610.02391.

[24] Llyod S. Shapley. *A Value for N-Person Games*. RAND Corporation, Santa Monica, 1952. doi:10.7249/P0295.

[25] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences, October 2019. URL http://arxiv.org/abs/1704.02685.

[26] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*, September 2014. URL http://arxiv.org/abs/1409.1556.

[27] Mukund Sundararajan and Amir Najmi. The many Shapley values for model explanation, February 2020. URL http://arxiv.org/abs/1908.08474.

[28] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks, June 2017. URL http://arxiv.org/abs/1703.01365.

[29] Hanxiao Tan. Maximum Entropy Baseline for Integrated Gradients, April 2022. URL `http://arxiv.org/abs/2204.05948`.

[30] D. J. Tolhurst, Y. Tadmor, and Tang Chao. Amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, 12(2):229–232, 1992. ISSN 02755408, 14751313. doi:10.1111/j.1475-1313.1992.tb00296.x.

[31] Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14(3):391–412, January 2003. ISSN 0954-898X, 1361-6536. doi:10.1088/0954-898X_14_3_302.

[32] Christopher Tyler, Lora Likova, Chien-Chung Chen, Leonid Kontsevich, Mark Schira, and Alex Wade. Extended Concepts of Occipital Retinotopy. *Current Medical Imaging Reviews*, 1(3):319–329, November 2005. ISSN 15734056. doi:10.2174/157340505774574772.

[33] Brian A Wandell, Alyssa A Brewer, and Robert F Dougherty. Visual field map clusters in human cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):693–707, April 2005. ISSN 0962-8436, 1471-2970. doi:10.1098/rstb.2005.1628.

[34] Brian A. Wandell, Serge O. Dumoulin, and Alyssa A. Brewer. Visual Field Maps in Human Cortex. *Neuron*, 56(2):366–383, October 2007. ISSN 08966273. doi:10.1016/j.neuron.2007.10.012.

[35] D. L. Yamins and J. J. DiCarlo. Using Goal-Driven Deep Learning Models to Understand Sensory Cortex. *Nature Neuroscience*, 19(3):356–365, February 2016.

A Appendix

A.1 Random baseline completeness

This is the development of the *completeness* axiom (Eq.2), employed with random baselines (Eq.1).

$$
\begin{aligned}
g_i(f, \mathbf{x}, \bar{\mathbf{x}}) &= \sum_{j=1}^{m} \mathbb{E}_{\bar{\mathbf{x}} \sim \mathcal{D}}[g_{i,j}(f, \mathbf{x}, \bar{\boldsymbol{x}}_k)] \\
&= \mathbb{E}_{\bar{\mathbf{x}} \sim \mathcal{D}}\big[\sum_{j=1}^{m} g_{i,j}(f, \mathbf{x}, \bar{\boldsymbol{x}}_k)\big] \\
&= \mathbb{E}_{\bar{\mathbf{x}} \sim \mathcal{D}}[f(\mathbf{x}) - f(\bar{\boldsymbol{x}}_k)] \\
&= f(\mathbf{x}) - \mathbb{E}_{\bar{\mathbf{x}} \sim \mathcal{D}}[f(\bar{\boldsymbol{x}}_k)] \\
&= f(\mathbf{x}) - \mu_{f(\bar{\mathbf{x}})}
\end{aligned}
\tag{A10}
$$

A.2 Implementation invariance

Our method inherits from IG the *implementation invariance* axiom, and we investigate the robustness of IGC results upon architectural changes for models with comparable performance. Unlike the model described in Subsection.4.3 (with results in Fig.6), the model of Fig.A7a does not have convolution layers (>97% accuracy), and the model of Fig.A7b has ReLU activation functions instead of Mish (>99% accuracy). The absence of convolution layers seems to allow slightly sharper local changes of IGC attributions, while ReLU activation functions let appear some checkerboard artifacts. However, in both cases, observed differences are minor and do not impact overall attribution patterns.

A.3 Baseline Shapley Correlation

Fig.A8 shows the results of a correlation-based dataset-wise attribution method like IGC, but built upon Baseline Shapley instead of Integrated Gradients. Baseline Shapley is slow to compute, so for practical reasons, we used only 10% of MNIST validation images, i.e. 1k images. Nonetheless, difference from IGC maps in Fig.6 are already negligible.
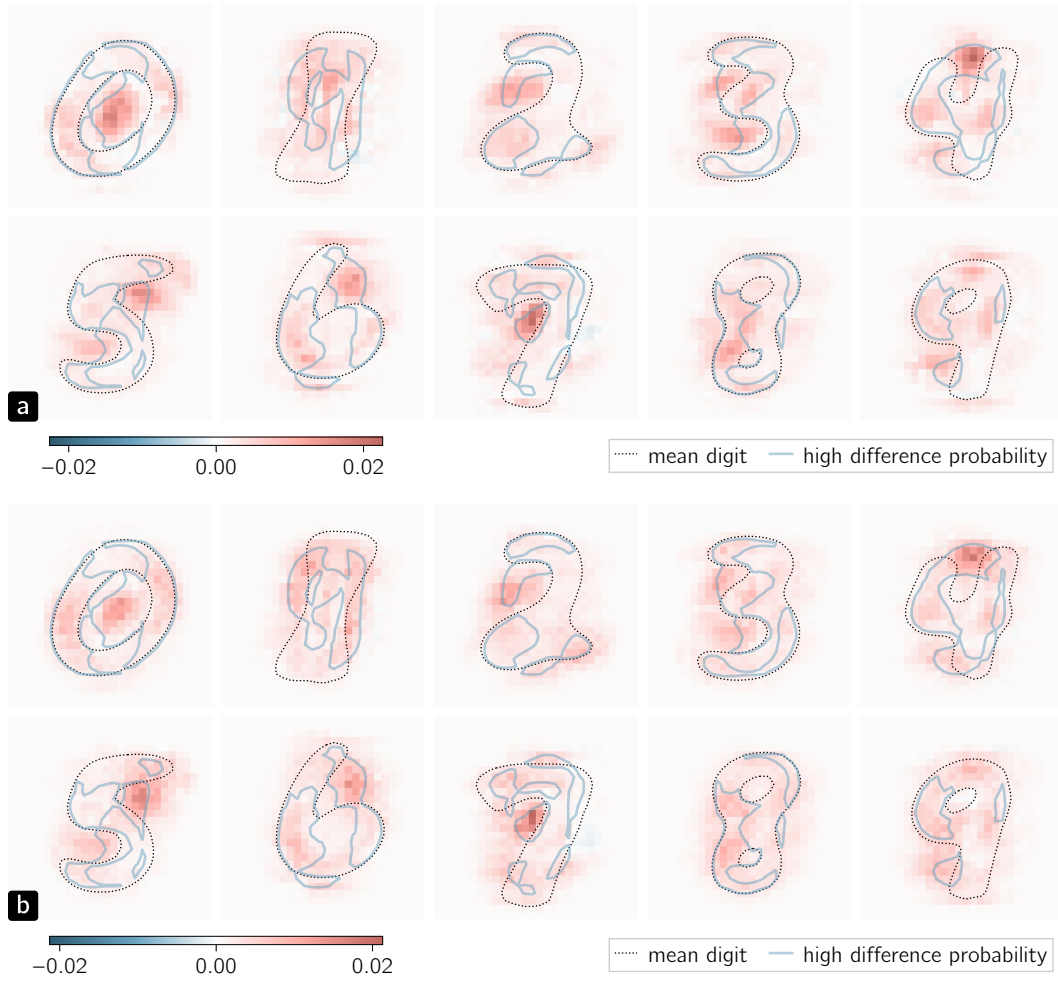
Figure A7: IGC maps for the MNIST dataset w.r.t. all possible digit classes. Employed models are a multilayer perceptron with Mish activation functions in panel **a**, and a convolutional model with ReLU activation functions in panel **b**. See caption of Fig.6 for details.
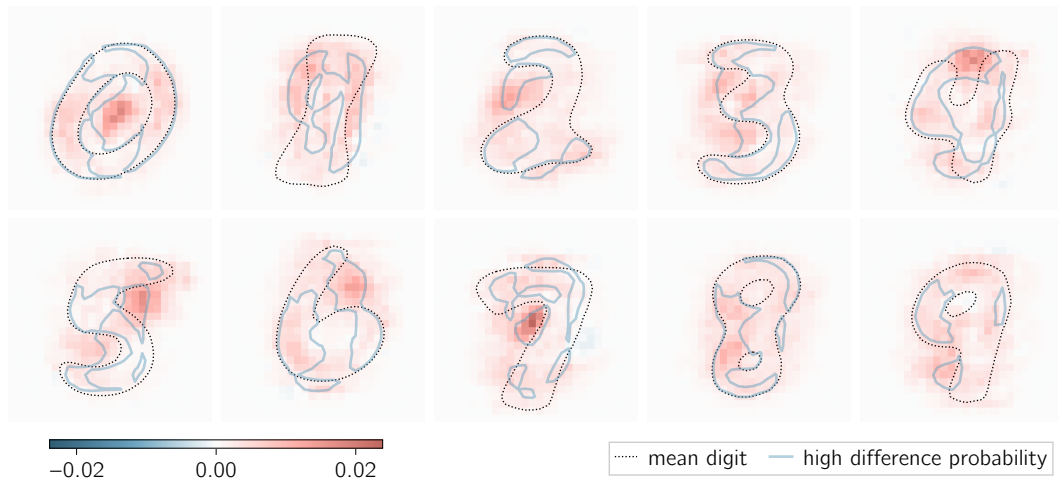

Figure A8: Baseline Shapley Correlation maps for the MNIST dataset w.r.t. all possible digit classes. See caption of Fig.6 for details.